LAMP-TR-112 June 2004
CAR-TR-996
CS-TR-4595
UMIACS-TR-2004-38

SYMBOLIC MT WITH STATISTICAL NLP COMPONENTS

Bonnie J. Dorr, Nizar Y. Habash, and Christof Monz

Institute for Advanced Computer Studies
University of Maryland
College Park, MD 20742-3275
{*bonnie,habash,christof*}*@umiacs.umd.edu*

Abstract

This reports provides an overview of the findings and software that have evolved from the ”Symbolic MT with Statistical NLP Components” project over the last year. We present the major goals that have been achieved and discuss some of the open issues that we intend to address in the near future. This report also contains some details on the usage of some software that has been implemented during the project.

Keywords: Machine Translation, Divergence Unraveling, Resource Projection

The support of this research by the Army Research Laboratory under contract ARL Cooperative Agreement DAAD190320020 is gratefully acknowledged.

Report Documentation Page

Form Approved
OMB No. 0704-0188

1. REPORT DATE	2. REPORT TYPE	3. DATES COVERED
JUN 2004		**00-06-2004 to 00-06-2004**

4. TITLE AND SUBTITLE	5a. CONTRACT NUMBER
Symbolic MT With Statistical NLP Components	5b. GRANT NUMBER
	5c. PROGRAM ELEMENT NUMBER
6. AUTHOR(S)	5d. PROJECT NUMBER
	5e. TASK NUMBER
	5f. WORK UNIT NUMBER
7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES) **Language and Media Processing Laboratory,Institute for Advanced Computer Studies,University of Maryland,College Park,MD,20742-3275**	8. PERFORMING ORGANIZATION REPORT NUMBER
9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)	10. SPONSOR/MONITOR'S ACRONYM(S)
	11. SPONSOR/MONITOR'S REPORT NUMBER(S)

12. DISTRIBUTION/AVAILABILITY STATEMENT
Approved for public release; distribution unlimited

13. SUPPLEMENTARY NOTES

14. ABSTRACT

15. SUBJECT TERMS

16. SECURITY CLASSIFICATION OF:			17. LIMITATION OF ABSTRACT	18. NUMBER OF PAGES	19a. NAME OF RESPONSIBLE PERSON
a. REPORT **unclassified**	b. ABSTRACT **unclassified**	c. THIS PAGE **unclassified**		**11**	

1 Participants

PI:

Bonnie Dorr, University of Maryland (UMCP), bonnie@umiacs.umd.edu

OTHER SENIOR PERSONNEL (Ph.D.):

Co-PIs

POSTDOCS:

Nizar Habash, Postdoctoral Researcher, University of Maryland, habash@umiacs.umd.edu

Christof Monz, Postdoctoral Researcher, University of Maryland, christof@umiacs.umd.edu

STUDENTS:

Necip Fazil Ayan, University of Maryland, nfa@umiacs.umd.edu

Nitin Madnani, University of Maryland, nmadnani@umiacs.umd.edu

2 COLLABORATIONS (BROADLY CONCEIVED)

1. Presentations by Bonnie Dorr to Georgetown on the use of linguistic information in hybrid statistical/symbolic tasks (summarization, machine translation, divergence unraveling).
2. Collaboration with Philip Resnik for the JHU/ONR MURI project.

3 PROJECT FINDINGS

1. Translation divergences are frequently occurring: Examination of the Spanish-English parallel corpus shows that divergences occur in 35% of the sentence pairs.
2. Unraveling the divergences with linguistically motivated universal rules results in improved word-level alignments: Experiments for Spanish-English alignments show statistically significant improvements of DUSTer compared to a state-of-the-art statistical aligner (GIZA++).
3. Generation-Heavy Machine Translation has a higher degree of robustness than a statistical translation system (IBM-4) and scores higher when the test set is not from the same genre the statistical system was trained on.

4 OPPORTUNITIES FOR TRAINING AND DEVELOPMENT (AT ALL GRADE LEVELS)

The improved word-alignment corpora can be used for Machine Translation in a number of ways:

1. Improved translation dictionary extraction
2. Improved statistical machine translation

In addition, the universal rules help to identify sentences that contain divergences, and this information can be used to exclude them from parallel texts that are used to train a statistical aligner. This results in less complex a corpus, on which standard statistical aligners can be trained. This is currently examined by Dr. Philip Resnik for Chinese-English translation in the context of the MURI project.

The GHMT system has recently been adapted to Chinese, and at this point we are also adapting it to Arabic. GHMT is also used for cross-lingual summarization, where summarization and translation are fully integrated.

- OUTREACH ACTIVITIES (DEFINED TO BE OUTSIDE OUR PROFESSIONAL COMMUNITIES)

5 PUBLICATIONS AND PRODUCTS

5.1 JOURNAL/CONFERENCE PUBLICATIONS

Bonnie Dorr, Necip Fazil Ayan, and Nizar Habash. ”Divergence Unraveling for Word Alignment of Parallel Corpora”. Submitted to *Natural Language Engineering*.

Bonnie Dorr, Necip Fazil Ayan, Nizar Habash, Nitin Madnani, and Rebecca Hwa, ”Rapid Porting of DUSTer to Hindi”, *ACM Transactions on Asian Language Information Processing (TALIP)*, 2:3, 2003.

Nizar Habash and Bonnie Dorr, ”CatVar: A Database of Categorial Variations for English”, in *Proceedings of the MT Summit*, New Orleans, LA, pp. 471–474, 2003.

Nizar Habash and Bonnie Dorr, ”A Categorial Variation Database for English”, *Proceedings of North American Association for Computational Linguistics*, Edmonton, Canada, pp. 96–102, 2003.

Nizar Habash. ”Matador: A large scale Spanish-English GHMT system”. *In Proceedings of the MT-Summit*, pages 149 156, 2003.

Habash, Nizar, Bonnie J. Dorr, and David Traum, ”Hybrid Natural Language Generation from Lexical Conceptual Structures”, *Machine Translation*, 18:2, 2003.

5.2 ONE-TIME PUBLICATIONS (INCLUDES BOOK CHAPTERS AND DISSERTATIONS)

Nizar Habash. *Generation-Heavy Hybrid Machine Translation.* PhD Thesis, University of Maryland, 2003.

Necip Fazil Ayan. *Injecting Linguistic Information to Improve Word Alignments for Statistical MT Systems.* PhD Research Proposal, University of Maryland, 2004.

6 OTHER PRODUCTS

1. DUSTer: A word-aligner that combines statistical information and linguistic rules for divergence unraveling. Download: `http://clipdemos.umiacs.umd.edu/duster/duster.tar.gz`

   Installation:

   ```
   gunzip duster.tar.gz
   tar -xf duster.tar
   ```

   documentation:

   `DUSTer-Package/docs/README`

2. CatVar: A Categorial Variation Database for English. CatVar is an extensive is an extensive database of morphological variation for English. CatVar is integrated into GHMT in order to increase the flexibility of the generation of the English translation. Online demo: `http://clipdemos.umiacs.umd.edu/catvar/`

3. Parallel corpora for Spanish-English and Hindi-English with DUSTer word-level alignments.

4. Generation Heavy Machine Translation system (GHMT). Currently, GHMT supports Spanish to English and Chinese to English translation. Download: `http://clipdemos.umiacs.umd.edu/ghmt/GHMT-PAK.tar.gz`

   Installation:

   ```
   gunzip GHMT-PAK.tar.gz
   tar -xf GHMT-PAK.tar
   ```

   documentation:

   GHMT: `GHMT-PAK/GHMT/install.readme`

7 CONTRIBUTIONS

7.1 CONTRIBUTIONS WITHIN THE DISCIPLINE

1. An extensive database of morphological variation for English.
2. Improved word mappings between Spanish and English and Hindi and English.
3. Robust Machine Translation system from Spanish to English and Chinese to English.

7.2 CONTRIBUTIONS TO OTHER DISCIPLINES (THIS IS NOT EXPECTED FROM ALL PROJECTS)

The project is relevant to the augmentation of capabilities useful for intelligence analysts, such as cross-lingual summarization and data mining.

In addition, the corpora with improved word-level alignments can be used for general resource projection from English onto the foreign part of a parallel corpus.

- CONTRIBUTIONS TO THE DEVELOPMENT OF HUMAN RESOURCES (SPECIFIC FOCUS ON RESEARCH OPPORTUNITIES, UNDERREPRESENTED GROUPS, EDUCATIONAL MATERIALS, AND MEMBERS OF THE PUBLIC)

7.3 CONTRIBUTIONS TO RESOURCES FOR RESEARCH

This work provides an integral part for many NLP applications that require cross-lingual processing or processing in a resource-poor foreign language.

7.4 CONTRIBUTIONS BEYOND SCIENCE AND ENGINEERING (THESE CAN BE SPECULATIVE)

The research carried out in this project contributes to the development of better-performing machine translation systems. The availability of high-performance MT has far-reaching consequences for society in general, as it facilitates laymen and professionals in accessing information that is authored in a language they do not understand.

8 PLANS FOR THE NEXT YEAR, IF CHANGED

Recently, we have designed a prototype that will allow us to use many different resources, such as statistical aligners, linguist rules, cognate lists, and dictionaries, and combine their partial evidence to yield more accurate word-level alignments.

Additionally, we will adapt our GHMT implementation to Arabic-English translation.

Our funding for this project ends in the summer of 2005. We will need additional funds for the 2 years after the project has expired to continue the high level of activity toward this effort that we have contributed over the last year. For this, we have one proposal currently under review:

> ”Divergence Resolution for Interlingual Variation Encoding (DRIVE)”, REFLEX submission, Broad Agency Announcement (BAA-04-01-FH), May 2004.

9 SPECIAL REPORTING REQUIREMENTS, IF ANY

None.

10 UNOBLIGATED FUNDS (ONLY IF OVER 20%)

N/A

11 SIGNIFICANT CHANGE IN USE OF HUMAN SUBJECTS

None.

A DUSTer

usage: end-to-end.pl <argument-file>

end-to-end.pl processes an entire corpus of two languages and projects dependency trees and alignments from one corpus to another, using DUSTer. (See README.overview for more information on what DUSTer is) The only argument t o the script is a file specifying paths for external scripts and programs, and t he arguments for DUSTer.

In order to run DUSTer correctly, please make sure that you put the following into your shell files (into your .tcshrc file or whatever you use):

setenv DUSTERPATH <duster-path> set path = ($path $DUSTERPATH/bin)

DUSTERPATH should be set to the directory that contains pos, rel, bin, epgen, lib, docs directories. Without these statements and the correct setting of DUSTERPATH, DUSTer will not work! (After installing DUSTer on your system, make sure that you set DUSTERPATH correctly)

EXAMPLE RUN:

In order to run DUSTer on a small English-Spanish (or English-Hindi) example, go to examples/Spanish (or Hindi) directory and run

end-to-end.pl end-to-end-arguments.sp

or

end-to-end.pl end-to-end-arguments.hin

If you get comb-aligned directory under DUSTer-Results, that means you have run DUSTer successfully on this small example.

IMPORTANT REQUIREMENTS:

1. To run runGIZA++.pl, you should define the correct path for GIZA++ and EGYPT package in your shell file (i.e., GIZAPATH and EGYPTPATH). For example,

setenv GIZAPATH /dfs/projects/clip-proj/IBM-MT/bin
set path = ($GIZAPATH $path)

setenv EGYPTPATH /fs/clip-archive/connor/Corpora/not-really-corpora/cvs-export
s/EGYPT/bin set path = ($EGYPTPATH $path)

Make sure you have these programs on your machines and set these environment variables correctly.

2. To run Python files, you should have a correct pointer to Python version 2.3, as discussed above.

ARGUMENTS TO END-TO-END.PL

The only argument to the script is a file specifying paths for external scripts and programs, and the arguments for DUSTer.

Here is an example argument file (which is under examples/Spanish/end-to-end-arguments.sp) Please, make sure that the argument file contains a value for each of these variables (You can simply copy this file and change the values in the second column appropriately)

Program Paths

EPEXTRACT	$DUSTERPATH/lib/epextract.py
COMBALIGN	$DUSTERPATH/lib/combalign.py
GIZA2INFERDEPGRAPH	$DUSTERPATH/utils/giza2infer.pl
RUN_GIZA	$DUSTERPATH/utils/runGIZA++.pl

Make sure you change this to the path for Python 2.3 or later versions on your machines. The following is just an example path and probably will not be in your system. PYTHON /usr/local/stow/Python-2.3.2/bin/python2.3

Arguments

DUSTER_CONFIGURATION_FILE	config.spanish
LEFT_LANGUAGE	English
RIGHT_LANGUAGE	Spanish
LEFT_CORPUS	dev.eng
RIGHT_CORPUS	dev.sp
DEPENDENCY_TREE_DIR	dev.ecollins.dep-dir
ALIGNMENT_FILES_DIR	dev.giza.alignments-dir
OUTPUT_DIR	DUSTer_Results

The first four variables in the argument file correspond to some external programs distributed with this package. You don't need to change those lines unless you move those scripts to other directories.

PYTHON variable should be set to the path for Python version 2.3 or later. Otherwise, the python scripts will not work.

The rest of the arguments depends on the corpus you are running DUSTer on. DUSTER_CONFIGURATION_FILE is the file where DUSTer-specific paths and variables are set. This must be done for each language pair once and then it can be use on every parallel corpus on those two languages. For Hindi and Spanish, example configuration files are provided in the example directory (config.spanish and config.hindi). For more information about this configuration file, please see README.auto-DUSTer.

LEFT_CORPUS refers to the English corpus and RIGHT_CORPUS refers to the foreign language corpus. These should contain one sentence per line and they should be parallel (i.e., the sentences on the same line are translations of each other)

DEPENDENCY_TREE_DIR is the directory that contains dependency trees for English sentences. Each file MUST be named as tree.n where n is the number of the sentence in the corpus (tree.1, tree.2, etc.). The directory name should include the name of the parser. (The current standard is ¡corpus-name¿.¡parser¿.dep-dir (for example, devdata.ecollins.dep-dir or devdata.eminipar.dep-dir). The name of the parser is important to locate the necessary mapping files. For the formatting of each file, please see READMe.auto-DUSTer or example trees under example directory.

ALIGNMENT_FILES_DIR is the directory that contains the initial alignment files. Each file MUST be named a align.n where n is the number of the sentences in the corpus (align.1, align.2, etc.) Each file starts with ”begin n” and ends with ”end”. In between, each English word is associated with a bunch of FL words. Please, see READMe.auto-DUSTer or example alignment files under example directory for further details.

OUTPUT_DIR is the directory where all output files will be written. The sentences for which DUSTer runs successfully will be placed under succ directory. Otherwise, the related files for that sentence will be under err directory. The final alignment files (i.e., DUSTer alignments) will be placed under succ/comb-aligned directory. Each output file in this directory is named as ehmap.n where n is the sentence number.

B GHMT

REQUIRED RESOURCES

1. LISP: International Allegro CL Enterprise Edition 6.0 (Franz Inc.)
2. Perl: v5.8.0
3. Connexor parser (English and Spanish) (from www.connexor.com) See instructions below on hooking up the connexor client to the rest of the system.
4. Nitrogen Morphology Support
   Nitrogen is available at: http://www.isi.edu/natural-language/projects/nitrogen/
   Specifically, the morphology files, nitro.english.morph.lisp nitro.morph.8.98.lisp nitromorph-8-98.lisp must be placed under $PACKAGE/EXERGE/SOURCE/oxyexerge/
5. Halogen Forest Ranker
   Halogen is available at: http://www.isi.edu/licensed-sw/halogen/ All code from the forest ranker should be installed under $PACKAGE/HALOGEN/ForestRanker

Make sure the variables in sysVars.cshrc are added to your .cshrc

The source files for the Exerge system are included in this package in addition to created images on Solaris. to remake these images, run $PACKAGE/ake-Exerge.sh

See a Sample run of Matador below.

CONNEXOR SPECIFIC INSTRUCTIONS

1. Contact www.connexor.com to obtain a license for English and Spanish parsers.
2. Update the host/port in the files fdges-client.pl (for Spanish) and fdgen-client.pl (for English). The current values should look like this for fdges-client.pl:
   $remote_host="cheesecake.umiacs.umd.edu"
   $remote_port="11720"
   and as follows for fdgen-client.pl
   $remote_host="cheesecake.umiacs.umd.edu"
   $remote_port="11721"

SAMPLE RUN

```
> matador.pl test out2 x params=params.matador.2
parameter params = params.matador.2  ... loading
Processing Batch #0
PARSING...
TRANSLATING...
reading /fs/clip-plus/habash/PACKAGE/TRANSLEX/Spanish-English/span-eng.tralex ... done
translating torotemp.dep ...
done
CONVERTING,EXPANDING...
/fs/clip-plus/habash/PACKAGE/EXERGE/corexerge.sh torotemp.trans.amr torotemp.out.amr
T NIL T T T 10 10 10 10 NIL 1 T NIL T
<Running CorExerge>
; Exiting Lisp
LINEARIZAING...
```

```
<Running OxyGen 2.0>
; Exiting Lisp
RANKING...
/fs/clip-plus/habash/PACKAGE/EXERGE/halogenize torotemp.out.gls torotemp.out.txt 6
/fs/clip-plus/habash/PACKAGE/HALOGEN/ForestRanker/news.binlm

<GLS-to-Forest Conversion> && <Running  HALOGEN>

/fs/clip-plus/habash/HALOGEN/ForestRanker/polishsen.pl
/fs/clip-plus/habash/PACKAGE/MATADOR/halolin-temp.sen0
> /fs/clip-plus/habash/PACKAGE/MATADOR/halolin-temp.sen
; cpu time (non-gc) 420 msec user, 10 msec system
; cpu time (gc)     70 msec user, 0 msec system
; cpu time (total)  490 msec user, 10 msec system
; real time  23,139 msec
; space allocation:
;  332,622 cons cells, 7,882,064 other bytes, 0 static bytes; Exiting Lisp
REPORTING...
done!
```